View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Xiamen University Institutional Repository

学校编码：10384　　分类号_____密级_____
学号：200340009　　UDC _____

厦门大学

硕　士　学　位　论　文

# 一种具有英汉辅助翻译功能的拼音输入法

# A Pinyin Input Method Editor with English-Chinese Aided Translation Function

李　栋
指导教师姓名：史晓东　教授

专　业　名　称：计算机应用技术
论文提交日期：2006 年 5 月
论文答辩时间：2006 年 6 月
学位授予日期：

答辩委员会主席：_____
评　　阅　　人：_____

**2006** 年 **5** 月

厦门大学博硕士论文摘要库

# 厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学博硕士论文摘要库

# 厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

厦门大学博硕士论文摘要库

# 摘 要

拼音输入法不仅具有易学易会的特点，而且拥有相当多的用户。目前市场上的拼音输入法一般是以词为单位进行输入的，具有拼音串可编辑和汉字串可编辑、混合输入、模糊输入、机器学习、特殊处理、自动转化（智能化）等特点，比较典型的有：微软拼音输入法、紫光输入法、拼音加加输入法等。

机器翻译是指通过计算机来实现不同自然语言之间的翻译,现正逐渐成为克服不同语言交流障碍的重要手段。目前基于统计的方法已经成为机器翻译研究领域的主要研究方向之一。统计机器翻译可以利用海量语料库直接训练得到机器翻译系统，具有无需人工干预、译文质量机械味道少等优点。

通常的拼音输入法重码率高，面向普通用户使用，而对专业从事翻译的人群来说，这些输入法则忽略了行业的特殊性和便利性。基于统计的机器翻译技术在整句或成段英文翻译效果方面差强人意。将拼音输入法和统计机器翻译技术结合起来，不仅可使拼音输入法具有辅助翻译功能，而且通过专业翻译人员和拼音输入法的交互，也可改善统计机器翻译中英文句子的翻译质量，产生符合专业翻译人员的要求的译文。

基于上述思想，本文开发了一种供专业翻译人员使用的交互式翻译工具——具有英汉辅助翻译功能的拼音输入法。该工具从专业翻译人员的翻译思路出发，将目前已有的拼音输入法、基于统计的机器翻译和 n 元组统计技术结合起来，从而减少了翻译人员的击键次数，节省了翻译时间。

本文的创新之处如下：

1、依据专业翻译人员的思路，提出了英汉双语句子中词汇模糊对齐模型，利用该模型可估计翻译人员当前预翻译单词的译文，排除单词翻译的歧义性。

2、以英汉双语句子中的词汇模糊对齐为基础，将汉字拼音输入法和基于统计的机器翻译技术结合起来，估算翻译人员想要的字词，供翻译人员确认选择，从而减少击键次数，产生高质量的译文。

**关键词：**拼音输入法；英汉双语句子中的词汇模糊对齐；n 元组统计模型；

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

# Abstract

PinYin Input Edit Method is easy to grasp, and has a great many customers in market. Nowadays, PinYin Input Edit Methods usually accept a word as input, and they can edit PinYin Clusters and Chinese Characters Clusters, also they are characteristic of mix input, blur input, machine learning, special processing, and automatic transferring (intelligence) etc. Typically, there are Microsoft PinYin Edit Method, ZiGuang Edit Method, and PinYin jiajia Edit Method and so forth.

Machine Translation implements the translation of different nature languages by computer, and gradually becomes an important means in overcoming the language barrier. Presently, statistic-based method is one of the major research fields in machine translation. Statistical machine translation (SMT) doesn't need human intervention, gets the machine translation system by direct training with the large corpora, and reduces the machine characteristics of translated articles.

Usually, PinYin Edit Methods have high-duplicate-code, and object to general customers, however, to professional translators, they neglects the particularity and particularity of the translation industry. Statistic-based machine translation method doesn't work excellently in translating a whole sentence or paragraph. While, combining the PinYin Edit Method and Statistical machine translation, PinYin Edit Method can assist translation, but also improve the translation quality of Chinese-English sentences with the interaction of professional translators and PinYin Edit Methods, generating the required translations.

Accordingly, this thesis develops an interactive translation tool objected to professional translators -- a PinYin Edit Method with English-Chinese aided translation function. This tool, from the point of view of professional translators, integrates current PinYin Edit Methods, SMT and n-gram technology, reduces the tapping times and saves translation time.

Innovation of this thesis:

1. Aiming at the professional translators, present a fuzzy alignment of vocabulary in English-Chinese bilingual sentences. This model can assess the current pre-translation, eliminate the ambiguity of words.

2. Based on the fuzzy alignment of vocabulary in bilingual sentences, combine the Chinese Character PinYin Edit Method and SMT, estimate the

厦门大学博硕士论文摘要库

words wanted by translators and offer the choices, which can lower the tapping times and generate required translation.

**Key words**：Pinyin IME；Fuzzy Alignment of Vocabulary in English-Chinese Bilingual Sentences；N-gram Statistical Method

厦门大学博硕士论文摘要库

# 目 录



厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

# Contents



厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

# 第 1 章　绪　论

## 1.1 课题的提出

语言是人类社会文化的主要载体和信息交流的重要工具。随着经济的全球化、计算机的普及和 Internet 的迅猛发展，世界上不同地区的经济文化联系日益密切，人们日常工作生活的信息化和国际化程度不断提高，语言的差异已成为信息交流中面临的严重障碍。将多种语言信息，特别是英文信息翻译成中文信息的工作量正在日益加大，这对翻译工作者来说不仅是一种挑战，而且也是一种沉重的负担。单纯依靠人工翻译，工作量大且速度有限，人们对实用的机器翻译系统的需求急剧增长。尽管经过无数机器翻译专家们的执著研究和不断探索，机器翻译无论在理论技术还是在实际应用方面都取得了长足的进步，但不可否认的是，现有机器翻译或辅助翻译系统的翻译质量还难以满足实用需要。所以很多专业翻译人员往往宁愿采用人工翻译的方式，而不愿采用现有的机器翻译或辅助翻译系统。在这种情况下，如何减轻翻译人员在使用计算机进行翻译及录入中的工作量就成为一个可研究的课题。基于此本文提出了一种具有英汉翻译辅助功能的拼音输入法。

## 1.2 课题的科学意义

具有英汉辅助翻译功能的拼音输入法主要涉及两方面的技术，即拼音输入法和基于统计的机器翻译技术（Statistics_based Machine Translation）。拼音输入法具有易学易会的特点，很容易让人接受，目前在市场上的用户占有率是相当大的，而且其中有许多种都具有词组学习记忆功能，比较典型的有：紫光输入法、拼音加加输入法、微软拼音输入法等。但这些输入法都是提供给一般用户使用的，具有普遍性，没有针对性。例如对专业从事翻译的人群来说，这些软件提供的功能并没有考虑到行业的特殊性和便利性。而具有英汉辅助翻译功能的拼音输入法是专为翻译人员设计的交互式翻译工具，以英文句子“I went to Italy last summer.”为例，当要翻译单词“summer”时，紫光输入法需要输入“xiatian”，微软拼音输入法需要输入“xiat”，才能将“夏天”显示在输入法备选列表的第一位，而使

厦门大学博硕士论文摘要库

用具有英汉翻译辅助功能的输入法时，用户只需输入“xt”，系统便会提示“夏天”及其它相关词组，并且将“夏天”这个提示放在第一个备选项上，这样便可以将用户敲击键盘的次数至少减少两次，节省了用户的翻译时间，提高翻译效率。设想如果是面对长篇累牍的英文文章，那么为专业翻译人员所节省的敲击键盘的次数累计起来会是相当可观的。

## 1.3 国内外研究现状

拼音输入法最大的缺点是重码率太高，经常需用户选择，这就加大用户敲击键盘的次数，浪费了用户的时间。为了解决同音严重的现象，不同系统采取了各种技术来提高输入速度和系统整体性能。目前比较流行的技术有：压缩编码长度、以词为单位进行输入、拼音串可编辑和汉字串可编辑、混合输入、模糊输入、机器学习、特殊处理、自动转化（智能化）等，特别是语句级输入技术的引入。然而语句级输入实质上是在系统内部对输入的分析和理解的过程，问题涉及到计算机科学、语言学、心理科学甚至哲学等领域的研究，因此问题的解决依赖于各学科的共同努力，才能达到满意的效果。

90 年代以来，机器翻译（Machine Translation，简称 MT）领域的方法基本上可以分为两类，即基于规则（Rule-Based）和基于语料库（Corpus-Based）的方法。基于规则的 MT 是传统方法，而基于语料库的 MT 则是 80 年代末以后发展起来的方法。基于规则的 MT 又可分为基于转换的方法和基于中间语言（Interlingua-Based）的方法，基于语料库的 MT 又可分为基于统计（Statistics-Based）的方法和基于实例（Example-Based）的方法。

统计机器翻译（Statistics-Based Machine Translation），又称为数据驱动（data-driven）的机器翻译，目前已逐渐成为国际上机器翻译研究的主流方法之一。其方法模型大致可以分为三种：第一种是基于平行概率语法的统计机器翻译方法，这种方法的主要代表有Alshawi的Head Transducer和吴德恺的SITG[1]（Stochastic Inversion Transduction Grammar）模型。第二种是基于信源信道模型的统计机器翻译方法[2,3]，它是目前最有影响的统计机器翻译方法，一般所说的统计机器翻译都是指这种方法。第三种方法是德国Och等人最近提出基于最大熵的统计机器翻译方法[4]。这种方法是比信源信道模型更一般化的一种方法。

厦门大学博硕士论文摘要库

交互式机器翻译是以牺牲全自动的翻译要求而获取较高质量译文的一种翻译方法。目前大部分系统的人工干预仅限于译前编辑、译后编辑，尤其是译后编辑。更为深入的人机交互式翻译研究追求的目标是允许用户在翻译的任何一个阶段都可以参与。这类研究可以根据人机交互发生的阶段分为：

1、交互式分析，用户帮助系统得出正确的源语言结构，尤其是复杂句子，对多义词进行排歧等。

2、交互式转换，用户参与选择与源语言结构等价的目标语言结构，排除不适当的转换。

3、交互式生成，用户协助产生流畅译文，用户在省略、指代、主题化方面对生成提供指导。

最早的交互式机器翻译（Interactive Machine Translation, IMT）工具是Kay[5]的“MIND system”。在这个工具中，用户的作用是通过回答有关源文本中单词认知、单词参照和前置短语关联等问题，来消除源文本中翻译的歧义性。后来的研究者，比如Blanchon[6]、Boitet[7]、 Brown和Nirenburg[8]、Maruyama和Watanabe[9]、Melby[10]、 Tomita[11]、Whitelock etal[12]、Zajac[13]等开发的系统也基本上沿用这种模式，他们的研究主要集中在提高消除歧义的效率和通过问题排序等技术来减少问题数量从而减轻用户的负担；根据源语言可供选择的解释找到更自然的表达方式；给用户提供多种翻译选择，而且把最可能的翻译放在所选择内容的首位。尽管这些研究者付出了巨大的努力，但 “回答问题式”的IMT系统仍旧相当耗时费力，因此这类系统主要应用在用户对目标语言的知识有限或对目标语言知之甚少的情况下。

Guy Lapalme和Philippe Langlais等于 1997－2005 年间开发一个专供翻译人员使用的交互式辅助翻译工具TransType[14]。TransType的新颖之处在于将目前的基于统计的机器翻译和输入法技术结合起来，旨在减少翻译人员的翻译工作量。目前已经有了TransType2 版本。TransType首先会扫描翻译人员预翻译的源语言文本，再根据翻译人员的键入的文本，给出后继翻译的提示，用户可以接受、修改这个翻译提示，也可以继续键入文本来忽略这个提示，系统则会不断地依据用户每次新的击键，重新计算然后修正对当前文本翻译的提示。总的来说TransType可以在以下三个方面给翻译人员的工作予以帮助：

1、当翻译人员在脑子里已经对要翻译的文本有合适的措词时，系统的提示

厦门大学博硕士论文摘要库

可以加快翻译人员的翻译速度；

2、当翻译人员在脑子里对要翻译的文本尚无合适的表达时，系统的提示可以充当一个建议，启发翻译工作者；

3、当翻译人员在翻译录入时出现拼写错误，系统的提示可以充当一个警告。

现在 TransType2 已经实现了三种语言 English ↔ Spanish、English ↔ French、English↔German 之间的相互辅助翻译。

## 1.4 本文的主要工作及创新之处

本文实现了一种供专业翻译人员使用的交互式翻译工具——具有英汉辅助翻译功能的拼音输入法。其主要工作集中在以下几点：

1、利用机器翻译工具包和分词工具实现了英文句子和单词的翻译及句子译文的分词。以此为基础，从专业翻译人员的思路出发，提出了英汉双语句子中的词汇模糊对齐策略，它是本文实现的立论基础。

2、研究了 n 元组统计语言模型理论基础——隐马尔可夫链及其相关技术，包括 n 元组模型的参数估计、模型压缩等，建立了《人民日报》语料库 1 元组、2 元组、3 元组的统计文件，并对 2、3 元组的统计结果建立了二级索引。以此为基础来排除在估算翻译人员心目中互译句对的词汇对齐模型时，汉语字词的重音现象对系统判断的干扰。

3、深入研究了 Windows 平台下的汉字输入法技术，其中分析了 Windows 系统下 IME 实现的基本原理、IME 的组成、输入法的体系结构、输入法设计中用到一些重要结构和消息及其工作的方式、IME 和不同应用系统之间的通信，实现了一种支持词组输入功能的拼音输入法。

4、设计了具有英汉辅助翻译功能拼音输入法的系统框架，实现了具有英汉辅助翻译功能的拼音输入法。并以手工方式对系统进行了评测。同时也对系统中存在的一些问题进行了分析说明。

本文的创新之处如下：

1、依据专业翻译人员的思路，提出了英汉双语句子中词汇模糊对齐模型，利用该模型可估计翻译人员当前预翻译单词的译文，排除单词翻译的歧义性。

2、以英汉双语句子中的词汇模糊对齐为基础，将汉字拼音输入法和基于统计的机器翻译技术结合起来，估计翻译人员想要的字词，供翻译人员确认选择，

厦门大学博硕士论文摘要库

从而减少击键次数，产生符合专业翻译人员要求的译文。

## 1.5 本文的组织结构

本文的主要内容分为 7 章：

第 1 章是绪论，介绍了本课题的立题依据、科学意义、国内外研究现状、主要工作、创新之处和本文的组织结构。

第 2 章是英汉双语句子中词汇模糊对齐，详细介绍了英汉双语句子中词汇模糊对齐策略的概念，实现基础和实现策略。

第 3 章是 n 元组统计语言模型，介绍了 n 元组统计语言模型的相关技术，对 1998 年《人民日报》1 月份的语料库进行了 n 元组统计和索引。

第 4 章是 Windows 平台下的汉字输入法技术，在分析了 IME 技术基础上，实现了一种 Windows 平台下汉字拼音输入法程序。

第 5 章是系统的设计和实现，首先设计了具有辅助翻译功能的拼音输入法的系统框架，接着实现了“把英文文本切分为句子模块”、“汉字注音模块”、“拼音输入法改造”模块；然后实现了英汉辅助翻译引擎程序和拼音输入法程序进程间的通信，最后详尽地描述了具有英汉辅助翻译功能的拼音输入法的整体运行机制。

第 6 章是系统评测，通过手工方式评测系统在句子翻译中的效果，并在击键次数方面与紫光拼音输入法和微软拼音输入法作了比较。实验结果表明具有英汉辅助翻译功能的拼音输入法对专业翻译人员是有帮助的。此外对实验中存在的问题进行了分析说明。

第 7 章是工作总结和展望，简单总结了目前所做的工作，并提出了进一步的研究设想。

## 1.6 小 结

本章首先介绍了具有英汉辅助翻译功能的拼音输入法的立题依据、科学意义、学术思想以及相关技术的国内外研究现状，其中着重介绍了交互式辅助翻译工具 TransType，最后介绍了本文的主要工作、创新之处和组织结构。

厦门大学博硕士论文摘要库

Degree papers are in the “Xiamen University Electronic Theses and Dissertations Database”. Full texts are available in the following ways:
1. If your library is a CALIS member libraries, please log on http://etd.calis.edu.cn/ and submit requests online, or consult the interlibrary loan department in your library.
2. For users of non-CALIS member libraries, please mail to etd@xmu.edu.cn for delivery details.

厦门大学博硕士论文摘要库